DOCUMENT RESUME

ED 027 540 AL 001 812

By-Robinson, Peter

Basic Factors in the Choice, Composition and Adaption of Second Language Tests.

Pub Date Mar 69

Note-11p.; Paper given at the third annual TESOL Convention, Chicago, Illinois, March 5-8, 1969.

EDRS Price MF-$0.25 HC-$0.65

Descriptors-Aptitude Tests, Diagnostic Tests, *English (Second Language), *Language Tests, *Second Language Learning, *Test Construction, *Test Selection

Identifiers-Classification Tests, *Evaluation Tests, Prediction Tests, Progress Tests

Generally speaking, the main purposes of second language tests are survey; didactic, psychological or sociological research; and evaluation, the latter being the concern of this paper. Evaluation tests measure the knowledge the learner has of the second or foreign languages, and may be subdivided into the following categories: (1) aptitude tests, which assess a person's capacity to learn another language; (2) diagnosis tests, which are either "inventory," and attempt to make a complete list of what the student knows in the various areas of the spoken and written language, or "error," which seek to identify and explain specific student mistakes; (3) classification tests, which divide students up into various levels of language competence for the purpose of forming homogeneous classes; (4) prediction tests, which are used to predict the student's handling of the second or foreign language in specific social and work situations where the second or foreign language is the only language used; and (5) progress tests, which try to measure the student's progress in a given program. Once the purpose of the test has been determined, the following stages fall into place--level, type, selection, form, gradation, order, number of items, administration of test, correction, and validation. These points are discussed in turn and are followed by a listing of recent writings on testing in second languages. (AMM)

ED061340

"Basic Factors in the Choice, Composition and Adaption of Second Language Tests"

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

By Peter Robinson
Head, TESOL Programme
Etudes Anglaises
Faculté des Lettres
Université Laval

(Paper given at the Third Annual
Convention of TESOL
Chicago, Illinois
March 5-8, 1969)

AL 001 812

The most important factors in the choice, composition and adaption of second language tests would seem to be the kind of test to use, oral comprehension test, reading comprehension test etc., and what language items the test should contain; but in fact these questions are of secondary importance and depend entirely on the purpose of the test and the kind of people the test sets out to measure.

Generally speaking, there are five main purposes: survey; didactic research; psychological research; sociological research; and finally evaluation with which this paper is concerned.

Survey tests are used to gather information about the second language competence of various ethnic groups in a particular country where more than one language is currently spoken, Belgium or Canada for example. Survey tests are of course not restricted to bilingual or multilingual situations, but can be applied in countries where one language is current to measure the foreign language competence of various groups, i.e., the oral French of secondary school-children in England.

Didactic research is concerned with the effectiveness of different teaching techniques, different manuals, programmes, audio-visual aids, even with the assessment of teaching competence. Tests are used to show that for example a particular teaching technique is more effective than another.

Psychological tests are concerned with the way a person learns another language and with the way the acquisition of the new language affects his mother tongue and his personality.

Sociological tests cover more or less the same area as psychological tests, but at the level of the group, not of the individual. The whole question of contact and conflict between groups speaking different languages is examined.

Evaluation tests subdivide into five main categories: aptitude, diagnosis, classification, prediction and progress. They are concerned with measuring the knowledge the learner has of the second or foreign language.

The first of these five categories of evaluation tests, aptitude, the object of which is to assess a person's capacity to learn another language, can be of great help to the teacher in giving him some idea of how far and how fast a certain prospective student may progress and what kind of help he may need. A distinction must be made here between the general aptitude test just described and the limited aptitude test which deals with the student's capacity to learn

a certain language. The latter test gives no indication of his aptitude at all, but simply identifies the kind of problems that the student will meet in learning that language.

Diagnosis tests break down into two sub-categories: inventory and error. The inventory category attempts to make as complete a list as possible of what the student knows in the various areas of the spoken and written language, while the error category seeks to identify and to explain specific student mistakes.

Classification tests divide students up into various levels of language competence for the purpose of forming homogeneous classes. These levels and their sub-divisions, beginning level 1, 2 and 3, intermediate level 1, 2 and 3, advanced level 1, 2 and 3 are arbitrary levels which are more or less clearly defined by the teacher and the programme director.

Prediction tests are used to predict the student's handling of the second or foreign language in specific social and work situations where the second or foreign language is the only language used. A good example of this kind of test is the admission test in English for foreign students applying for admission to an English-speaking university. The test selects a certain number of students who are thereby supposed to have the minimum competence in English required to begin their studies at the university.

Progress tests are tests that try to measure the student's progress in a given programme. There are two kinds of progress tests, the overall progress test and the interim progress test. The former measures the student's overall progress from the beginning to the end of the course, whereas the latter deals with the extent to which the student has learnt the material of one or more lessons.

Once the purpose of the test has been determined, the following stages fall into place: level; type; selection; form; gradation; order; number of items; administration of test; correction; and validation.

After purpose, level is the most critical stage, since it determines the type of test, the language items to be included in the test and the form the test will take. Level is simply the amount of English the test assumes that the student should know to meet the requirements of one or more situations. The following two examples, admission tests for foreign students applying for admission to an English-speaking university and progress tests in a given course will illustrate

what level means in practice.

In the first example, admission tests, the level is defined in terms of the following situations: attendance at lectures and seminars; amount of reading required; number of written assignments. The level will be the minimum amount of English required to function efficiently and adequately in those situations.

As regards the second example, progress tests, the level for the interim tests is the language content of the manual used in class, while the level for the overall tests is what the teacher and the programme director think the beginning, intermediate and advanced student should know.

The type of test to be used is entirely dependent upon the level. For university admission tests, oral comprehension, oral expression, reading comprehension, and composition tests cover the language skills in which the foreign student must possess a certain minimum competence in order to carry out his studies. In actual practice, only oral comprehension and reading comprehension tests are used, as it is extremely difficult to make a rapid, consistent assessment of the student's ability to write and to speak. As regards progress tests, the type of tests will be determined by what has been taught in class.

Selection, namely what to include in the test, is directly related to the level. In the case of university admission tests, selection is made in a series of stages: the first matter to settle is whether to select material from the undergraduate or graduate levels; the next question is to decide which lectures and seminars to record; then a list of vocabulary, grammar and phonetic items is drawn up from the recordings; next, a certain number of these items in a certain proportion are selected for inclusion in the test; finally, of the items selected for inclusion in the test a certain number are chosen to directly assess the student's competence. The procedure is obviously not so lengthy and complex as regards progress tests.

As regards the form of the test, two basic decisions have to be made: objective or non-objective form for the student's answers; particular variant of a test type, i.e., vocabulary, grammar, phonetic, or semantic oral comprehension test, any one or any combination of the above.

Objective test or form is a misnomer, as it gives the misleading impression that the test so described is an independent, detached, eminently reliable, scientific evaluation. In fact the objective test is not intrinsicly more

reliable than the non-objective type. The difference between the two is no more than a question of procedure. In the non-objective form, the student, in response to a series of questions, makes a free, active use of the second language, whereas, in the objective form, the answer is already given, and all he has to do is indicate by a mark which answer is more appropriate out of the four answers that appear with each question.

The characteristic feature therefore of the objective form is its limitation of the student's participation and choice to selecting the right answer out of four given possible answers.

In many cases, it is really only a choice between two possible answers, as the other two are so obviously wrong for the intermediate student that he can easily narrow the choice down to two, and thereby have a 50% chance of selecting the right answer by a simple guess. This can largely defeat the purpose of having four answers per question to reduce the chance factor; and make interpretation of the results extremely hazardous.

While some students are helped towards the right answer, other students' attention is distracted: they concentrate on the irrelevant answers and end up either by selecting a wrong one or by wasting too much time in finding the right one. They either do not finish the test or have to rush through certain parts of the test in order to complete it. Once again interpretation of the results is extremely hazardous.

The objective form, or multiple choice as it is sometimes called, with its four answers, one right, the other three completely wrong, does not discriminate between different levels of language competence. The student is not faced with a real choice, between four truly possible solutions, which, considered separately, are all equally correct, but which considered together, sort themselves out in order of probability as right solutions. When the choice is not real, when the possible answers are not scored according to their degree of appropriateness as the right solution, the test may fail to distinguish between the student who knows nothing, who knows something and who knows a great deal.

Choice between the correct answer and typical errors made by the student is a valid procedure for groups who have a known, particular error pattern. But obviously Spaniards and Frenchmen do not make the same kind of errors in learning

English, and a test effective with the Spanish group would be useless with the French, or any other different national group, or with a group made up of people of different nationalities, as is the case with university admission tests.

The irrelevant alternative answers of the objective test can take on a surprising relevance for particular national groups. The following example taken from a vocabulary test given to 618 French Canadian first year education students at Laval in 1968 illustrates this well. It is a question of choosing the right synonym for revise.

change	33%
see	43%
paint	2%
learn	22%

It is clear that one distractor was useless (paint) and this narrowed the choice down to three. The preference for see arises probably out of association with reviser in the mother tongue, which, unlike its English cognate, does not have the meaning of to change, but only to look at again with the possibility of modification. There is also in French close association both in usage and in origin between revoir and reviser. Learn may come from association with revision exercises, or more simply from students reading too much into the question, or even from students being unable to make up their mind between change and see.

Perhaps, the most pertinent criticism that could be made against the multiple choice objective form is that it attempts to evaluate the student's competence in a particular language skill in such a passive way, and on his performance in a very limited area covered by a very small number of questions.

As the student's language competence is assessed within the very limited range of a determined number of questions, 30-100 normally as regards any language skill, it is highly important that the range covered by the questions reflects as accurately as possible the situation in which the student uses the language. The particular variant of a type of test has to be chosen with this in mind. It would seem as regards university admission tests that oral comprehension tests that concern themselves with the student's ability to distinguish between certain sounds, to recognize certain grammatical forms, to recall names and numbers, to know the meaning of individual words, rather than to grasp the general meaning of one or two sentences, would be trying to predict the student's performance in

the lecture by insignificant and inappropriate criteria. A student's ability to consistently distinguish between b and p may not be crucial to his understanding of a lecture.

Gradation, the grading of the difficulty of the questions in the test, and order, the sequence in which those questions appear, is not a pure linguistic exercise. Gradation and order are also determined by the make up of the group and by the situation in which the group has and/or will use the second language. There is no standard system of gradation applicable to all groups and all tests, but simply one which is valid for a particular group. The gradation for an oral comprehension test for absolute beginners who have done 50 hours of English will evidently not be the same for an oral comprehension test administered to foreign students applying for admission to an English-speaking university.

Involved with gradation and at the same time with administration, namely, the conditions under which the students take the test, are a series of factors. The first series of factors, for want of a better term, can be designated as presentation factors, that is the way the content of the test is presented to the student, for example, whether the test content is presented orally or in a written form; if oral, whether a tape recorder is used; if oral, whether in the form of a dialogue; if a dialogue, duration and number of dialogues etc. ... The second group of factors can be called student participation factors, namely, the way the student indicates his answers to the test, for example, in writing or orally; if in writing, whether he writes complete sentences, whether he fills in missing words, whether he enters a mark in a box, etc. ... Lastly, there are what might be called locale factors, that is the kind of place in which the test is given.

Correction is more than just tabulating the scores. Correction is the quantitative assessment of the importance the author of the test attributes to each question and to each answer.

Standardization or normalization is the final and most critical stage in the composing of tests, since the whole usefulness of the test is assessed. Unfortunately, a statistically satisfying picture of scores plotted evenly along a normal curve is no guarantee of the test's linguistic usefulness. The score distribution is purely a result of the composition of the group, and varies from group to group. In fact a normalized test is no more than a test which produces the same

results with similar groups; whether these results mean anything linguistically is another matter.

While statistical profiles of tests are in no way an indication as to the test's worth as a test, they are essential in providing the necessary data on which to base the evaluation of the test's fulfilment of its goal.

For objective tests, and this holds good for the non-objective type, normalization procedure is basically a detailed, statistical analysis of the students'answers. The first analysis with the whole group involves noting down the number and percentage of students that chose each alternative answer, and then lists showing the different selections are drawn up. The second analysis is identical to the first, but this time the group is no longer treated as a whole but it is divided up into three or four sub-groups according to the mark in the test, for example, students with a mark between 0 and 50 are classified as weak; those with a mark between 50 and 80 are intermediate; and those between 80 and 100 are strong. The purpose of such an arbitrary division is to see whether the choice of each group follows a consistent pattern and whether the pattern differs considerably from the pattern for the whole group. The third analysis is a detailed comparison between the performance of similar groups on the same test. In this way it is possible to single out those questions which need to be revised: questions may be too easy even for the beginning group, while, in another question, even the strong group may find it too difficult; or again certain questions will appear to be ambiguous, as, each time the test is taken, similar groups of students vary considerably in their answers, while being consistent with other questions.

Statistics provide the means to isolate and measure the variations in the students' choice of answers. However, it is up to the author of the test, the linguist, the teacher, the programme director to explain these variations.

The writing of a test and the use of a test require that a certain, fundamental procedure be followed in order that useful results be achieved: purpose of test; type of group to be tested; level of English of group to be tested; type of evaluation test to be used; language skills to be tested; selection of test content; form of test; gradation; order; administration; correction; normalization. It has been clear that great care has to be exercised with the objective form, as it may produce a test that means nothing. One danger is that it may distract the student's

attention from the essential point; another danger is that it may make it easier for some students to guess the right answer; another point is that the student does not exercise any real choice; even a greater drawback, and perhaps the most important one is that the student's active participation is zero and that his use of the language is neither seen nor heard; and finally his use of the language is surmised from such little evidence.

Of course this does not mean that the objective form always produces unreliable results, but simply there are built-in defects. The following variants in the objective form may go some way in dealing with the inherent problems:

1) of the four alternatives all are possible, but one is clearly the best.
2) as above, but the other three alternatives are scored according to their degree of possibility.
3) only two alternatives are offered with the following modification -

    1- both are right.
    2- both are wrong.
    3- only the first one is right.
    4- only the second one is right.

4) the right answer is contrasted with typical errors made by a known group of students.

Beyond the special problems posed by the objective form are the basic questions of how to realize these simple purposes: know the group which is going to be tested; know the language skill which is to be measured; know the situation in which the language skill is to be used; and know the test which best and most quickly suits the group, the language skill and the situation. The realization of these apparently simple goals is made all the harder by the fact that each test is only valid for the group it was designed to measure, and that all levels of language competence are relative, arbitrarily fixed to meet some situation by the teacher, the programme director, the university admissions board etc. Consequently, the scores are not absolute and have no meaning outside the context for which they were made. Tests could only lose their arbitrary, relative quality, if it were possible to define, and this is rather utopian, what the language competence of the average mother tongue speaker was; and then the second language student's use of the language could be measured against the fixed standard of the native speaker.

But, even if this utopian venture were possible with the help of all the socio-linguists, it must be remembered that the mother tongue speaker is not equally

competent in all situations. Hence it follows that the student's second language competence would be assessed in terms of the mother speaker's proficiency in only one situation, and his performance in other situations would have to be inferred from the mother tongue speaker's performance in those situations.

The question must now be raised whether it is in fact possible to have a general proficiency test which is valid for all groups, whatever the situation where the second language is used, whatever the social cultural background of the student may be.

Some tests appear to be effective for a fair proportion of students in a given situation, though the social cultural background of the students is considerably varied. A good example of this is provided by some university admission tests in English for foreign students. Leaving aside the question whether the students so selected do well in their studies and are better than those not admitted, and accepting the premise that the tests are effective, it remains to be shown what it is that makes the tests so effective.

* * * * * * * * * * * * * * * * * * *

Bibliography

Frederick Barton Davis — Educational measurements and their interpretation. Belmont, California: Wadsworth, 1964, 422 p.

Robert L. Ebel — Measuring educational achievement. Englewood Cliffs, New Jersey: Prentice-Hall, 1965, 481 p.

Robert Lado — Language testing. London: Longmans, 1961, 389 p.

George Perren — Testing ability in English as a second language. Part I. Problems. English Language Teaching, 21, 2(1967), p. 99-106.

Testing ability in English as a second language. Part 2. Techniques. English Language Teaching, 21, 3(1967), p. 197-202.

Testing ability in English as a second language. Part 3. Spoken language. English Language Teaching, 22, I(1967), p. 22-29.

G.D. Pickett — A comparison of translation and blank-filling as testing techniques. English Language Teaching, 23, 1(1968), p. 21-26.

Henri Piéron — Examens et docimologie. Paris: Presses universitaires de France, 1963, 190 p.

Theodore H. Plaister — Testing aural comprehension: a culture fair approach. TESOL Quarterly, 1, 3(1967), p. 17-19.

André Rey — Connaissance de l'individu par les tests. Bruxelles: Charles Dessart, 1966, 224 p.

Bernard Spolsky — Language testing-the problem of validation. TESOL Quarterly, 2, 2(1968), p. 88-94.

Anne O. Stemmler — The LCT, language-cognition test (research edition) - a test for educationally disadvantaged school beginners. TESOL Quarterly, 1, 4(1967), p. 35-43.

John A. Upshur & Julia Fata (Eds.) — Problems in foreign language testing. Proceedings of a conference held at the University of Michigan, September 1967. Language Learning, Special Issue, No. 3, August 1968.

John A. Upshur — Testing foreign-language function in children. TESOL Quarterly, 1, 4(1967), p. 31-34.

Rebecca Valette — Modern language testing: a handbook. New York: Harcourt, Brace & World, 1967, 200 p.